4-25-2016

The Microbiome of the Built Environment: A Large-Scale Citizen Science Genomics Project

Christopher Garrett Wilson

The Microbiome of the Built Environment: A Large-Scale Citizen Science Genomics Project

by

Christopher Garrett Wilson

An Honors Capstone
submitted in partial fulfillment of the requirements
for the Honors Diploma
to

The Honors College

of

The University of Alabama in Huntsville

April 27, 2016

Honors Capstone Director: Shawn Levy, Ph.D.
Director of Genomic Services Lab, HudsonAlpha Institute for Biotechnology

Student: _____ (Signature) Date: 4/25/16

Instructor: _____ (Signature) Date: 4/25/16

Department Chair: _____ (Signature) Date: 4/29/16

Honors College Dean: _____ (Signature) Date: 5/9/16

HONORS COLLEGE
THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

Honors College
Frank Franz Hall
+1 (256) 824-6450 (voice)
+1 (256) 824-7339 (fax)
honors@uah.edu

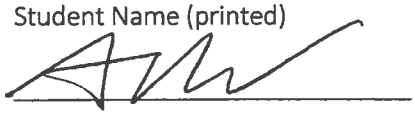
Honors Thesis Copyright Permission

This form must be signed by the student and submitted as a bound part of the thesis.

In presenting this thesis in partial fulfillment of the requirements for Honors Diploma or Certificate from The University of Alabama in Huntsville, I agree that the Library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by my advisor or, in his/her absence, by the Chair of the Department, Director of the Program, or the Dean of the Honors College. It is also understood that due recognition shall be given to me and to The University of Alabama in Huntsville in any scholarly use which may be made of any material in this thesis.

C. Garrett Wilson

Student Name (printed)



Student Signature

5/4/16

Date

The Microbiome of the Built Environment: A Large-Scale Citizen Science Genomics Project

Garrett Wilson,[1,2] Dan Dorset,[1] Melanie Robinson,[1] Shawn Levy[1]

[1]Genomic Services Lab, HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA; [2]Department of Biological Sciences, University of Alabama in Huntsville, Huntsville, Alabama 35899, USA

ABSTRACT

The microbiome represents the diversity of the microorganisms present in an environment, and the human microbiome has been increasingly recognized as an integral component of human health and disease. Thus, it is paramount to understand bacterial, viral, and metagenomic sources and distributions and how humans may interact with or acquire new commensal species or dangerous pathogens. The ultimate goal of metagenomics is to understand the structure and function of microbial communities, and there has been much interest generated from recent city-scale metagenomics studies of the built environment, such as Afshinnekoo et al. (2015). The metagenomic distribution of taxa from highly trafficked surfaces in a 9-12 grade high school has not yet been reported. James Clemens High School (JCHS) in Madison, AL, is an ideal candidate for this kind of study due to its relative new age as a built structure and its relative isolation in the community. The total enrollment in 2013 was 1,093

students (64% white, 26% African American, 7% Asian, and 4% Hispanic). We sought to characterize the JCHS metagenome by surveying the genetic material of the microorganisms and other DNA present in and around JCHS, with a focus on highly trafficked public areas (e.g. locker rooms, hallways, door handles, cafeteria, etc.). We envision this as a first step toward identifying potential pathogens, protecting the health of students, and providing a new layer of baseline molecular data that can be used by the school to create a "smart school," i.e., one that uses high-dimensional data to improve school planning, management of the built environment, and human health. To describe, characterize, and track the microbiome and metagenome of JCHS, we used next-generation DNA sequencing to profile the organisms present in our samples. This data establishes a school-scale, baseline metagenomic DNA profile.

INTRODUCTION

People spend more than 90% of their time indoors, where we breathe in and come in contact with trillions of microorganisms. Our homes, workplaces, hospitals and schools are quite literally complex ecosystems filled with a variety of microbes (Mason et al.). Given the amount of time we spend indoors it's important to understand what is living in these environments, how these microorganisms interact, what the potential implications are for human health, both positive and negative. The microbes that reside on the surface of the human body alone outnumber human cells by a factor of 10. The genomes of members of our indigenous microbial communities (the human metagenome) contain thousands of times more

genes than the human genome (Gill et al.). Microbial communities also inhabit the mouth, skin, and respiratory and female reproductive tracts. The compositions of these communities change over time. Understanding how microbial community structure affects human health and disease may contribute to better diagnosis, prevention, and treatment of disease. We may also learn things that could influence building construction practices and inform other industrial processes.

Today, architects and biologists are working together in ways previously unheard of to think about buildings using an ecosystem framework. This in and of itself is an advance in scientific inquiry. And while it would be a mistake to try and predict with certainty what will be discovered on this scientific frontier, we are building a body of research: a forecasting tool that will tell us what life forms are there based on other environmental variables.

This analysis of the built environment will identify and categorize what is living in the built environment, where we spend most of our time; it will describe how the life forms and their various communities interact among themselves and with humans. This is scientific inquiry in its purest sense. While it is hard to say what we will find, the opportunity exists to promote the growth of some species and inhibit the growth of others – just like we do with wildlife in national parks. In this case, however, there is the potential to impact human health. Metagenomic analyses of built environments provides us with high- dimensional data that could be used to improve city planning, management, and human health. We can establish baseline profiles across public spaces, identify potential bio-threats, and provide an additional level of data that can be used by city health officials and physicians.

One approach that has contributed greatly to understanding all organisms is genomics -- learning about the evolution and capabilities of organisms by deciphering the sequence of their DNA. Genomics has also greatly advanced microbiology, but, like pure culture, traditional genomics is limited in its ability to elucidate the dynamics of microbial communities (Board on Life Sciences). The decline in the cost of sequencing has made it possible to generate genomic sequences for a great variety of organisms (Huson et al.). Microbial sequences have since appeared at an exponentially increasing rate, with particular emphasis on pathogenic bacteria and eukaryotes -- such as the causative agents of plague, anthrax, tuberculosis, Lyme disease, cadidiasis, malaria, and sleeping sickness (Board on Life Sciences). We have made strides in improving the accuracy of matched reads with the introduction of whole-genome next-generation sequencing in metagenomics (Qin et al.).

The Genomic Services Lab partnered with the Mason Lab at Weill Cornell Medical College in New York City to survey the microbiome on a city-scale. Researchers in the Mason Lab swabbed every subway station in the city. The samples were sent to the GSL for processing and sequencing. The data we provided was used by researchers in the Mason Lab to create a heat map of microbes present in the various subway stations. This "PathoMap" is available for use online and all data has been released for further validation.

New York City is not the only city in the world that could benefit from a systematic, longitudinal metagenomic profile. Currently, built environment studies represent approximately 2% of all microbiome research (Stulberg et al.) The metagenomic distribution of taxa from highly trafficked surfaces in a high school has not yet been reported. James Clemens High School in Madison, AL, is an ideal candidate for this kind of study due to its relative new

age as a built structure and its relative isolation in the community. The total enrollment was 1,093 students in 2013, but has increased since this data was collected. We sought to characterize the JCHS metagenome by surveying the genetic material of the microorganisms and other DNA present in and around JCHS, with a focus on highly trafficked public areas (e.g. locker rooms, hallways, door handles, cafeteria, etc.) We envision this as a first step toward identifying potential pathogens, protecting the health of students, and providing a new layer of baseline molecular data to improve school planning, management of the built environment, and human health. To describe, characterize, and track the microbiome and metagenome of JCHS, we used next-generation sequencing to profile the organisms present in our samples.

MATERIALS AND METHODS

Sample collection

The Educational Outreach team at HudsonAlpha partnered with AP Biology students at JCHS to conduct environmental sampling. Samples were collected using Copan Liquid Amies Elution Swab 481C, a nylon-flocked swab with a 1 ml transport medium. The transport medium maintains a pH of 7.0 ± 0.5 and consists of sodium chloride, potassium chloride, calcium chloride, magnesium chloride, monopotassium phosphate, disodium phosphate, sodium thioglycollate, and distilled water (Amies, 1967). After a surface was sampled, the swab was immediately placed into the collection tube, coming into contact with the transport medium. Samples were then stored in a -4°C freezer once returned to the laboratory.

DNA extraction

Once samples were brought to room temperature, DNA was extracted using the MoBio Powersoil DNA isolation kit (as seen in Qin et al., 2010). Using the reagents from the kit, the cells in the sample were lysed, freeing the DNA and other contents. The other inorganic material was precipitated out. Using a concentrated salt solution, the DNA readily bound to the silica membrane of the kit's spin filters. An ethanol wash helps further clean and purify the DNA. Following the MoBio protocol, the 50 ml eluent was further purified by introducing 100 ml (2:1 ratio) of magnetic beads. Samples were then left to incubate at 25°C for 15 minutes and placed on an Invitrogen magnetic separation rack (MagnaRack) for 5 minutes. The DNA bound to the beads, and the supernatant was discarded. While the tubes were on the MagnaRack, 700

mL of 80% ethanol was added to the beads to wash off any remaining impurities. The ethanol was removed, and beads were left to dry. Finally, 10 mL of an elution buffer was added to purify the DNA, and 9 mL of the eluent was removed with 1 mL going toward QuBit quantification. Using a Qubit 2.0 fluorometer and the high-sensitivity kit (DNA HS standards, dsDNA HS buffer, and HS dye), we quantified the DNA in each sample. The parameters of the QuBit were set for ng/ml, and the value from the device was then multiplied by 8 mL for the total yield of the sample in ng.

Informatics

The SURPI metagenomics pipeline application was used in "comprehensive mode" to evaluate the unaligned FASTQ data produced by the sequencer (Naccache et al.). Briefly, the SURPI pipeline performs quality and adaptor trimming of reads using cutadapt. Trimmed reads are further filtered using SNAP to remove those reads which map to the human genome. The remaining reads were aligned using SNAP to the comprehensive NCBI nt DB. Of the various output formats provided by SURPI, the SAM files and taxonomic match tables were used for downstream processing. A custom script was written to traverse the taxonomic match tables for each sample and, for each sample, list the top 20 organism matches within several broad taxonomic category: viruses, bacteria, primates, non-primate mammals, non-mammal chordates, and non-chordate eukaryotes. The Entrez Taxonomic Database API was used to find the common name for organisms based on the taxonomic name, and if a common name was found it was included in the list.

Virulence marker analysis

A FASTA file containing pathogen-associated sequences was obtained from VibrioBase. The SAM file produced by SURPI was parsed, and those sequences aligning to bacterial genomes were extracted and aligned against the VibrioBase FASTA file using BWA (Li et al.). The SURPI organism and the best-match virulence marker were extracted for each aligned read in the SAM data. Those organisms which had at least 10 associated mappings to a single virulence DB contig were considered interesting. Constructed a list of the names of the virulence database records for these hits and pulled any nucleotide sequences aligning to those names from the SAM data. Created a FASTA file which specified the virulence database record name, followed by all of the nucleotide sequences which aligned to that virulence record. The FASTA was then BLASTed against the nt database, and wrote a simple XML parser that found the BLAST record associated with the top hit for each sequence (Wolfsberg et al.). Grouped and counted the times that each BLAST record was the top hit for a given virulence database match and wrote that info to a text file.

RESULTS

Baseline microbiome composition

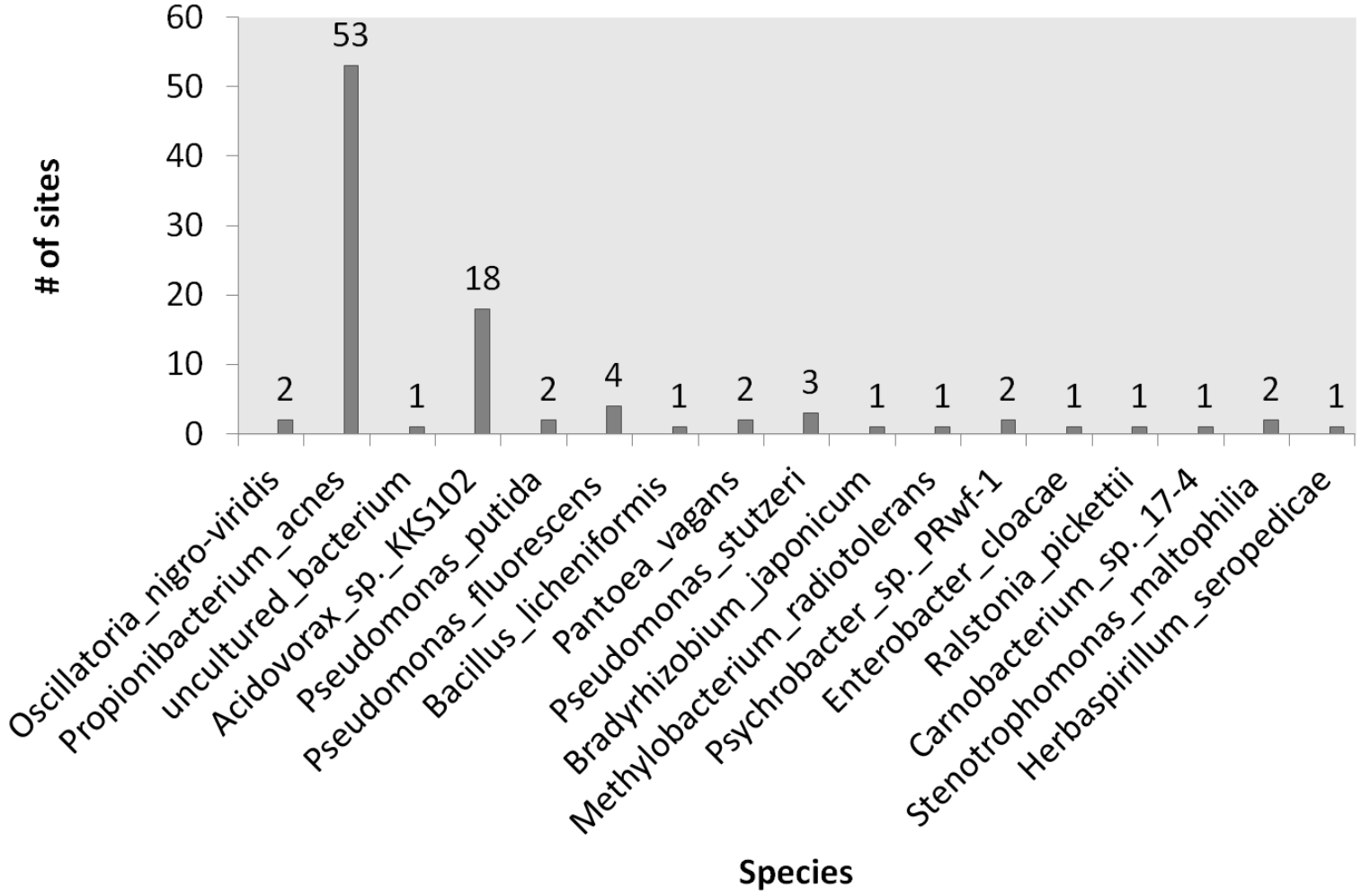


Fig. 1: Relative abundance of bacteria based on species presence at sampling locations. Microbiome composition from August 2014 sample collection. *P. acnes* and *Acidovorax* species are major contributors to the baseline composition.

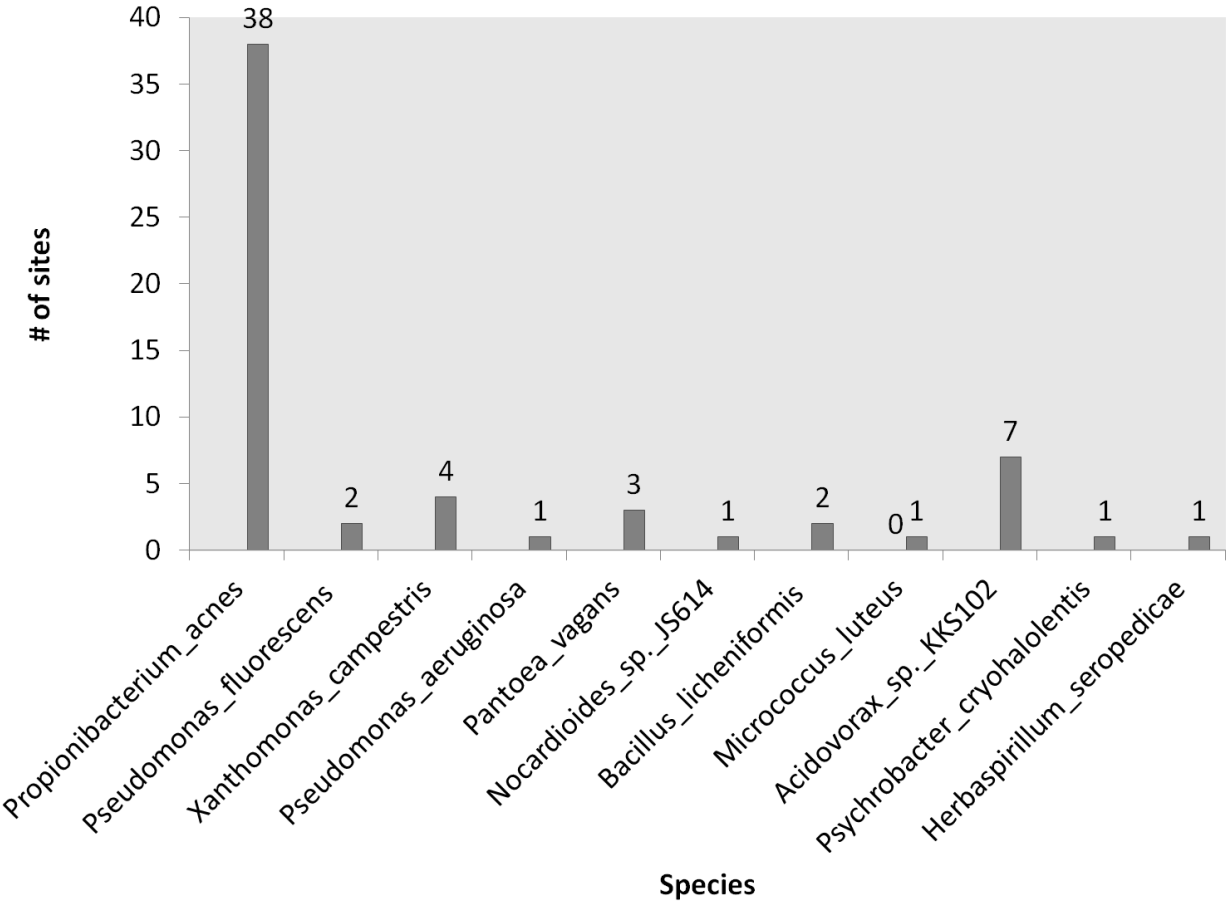


Fig. 2: Microbiome composition from September 2014 sample collection. *P. acnes* and *Acidovorax* are major contributors to the baseline composition.

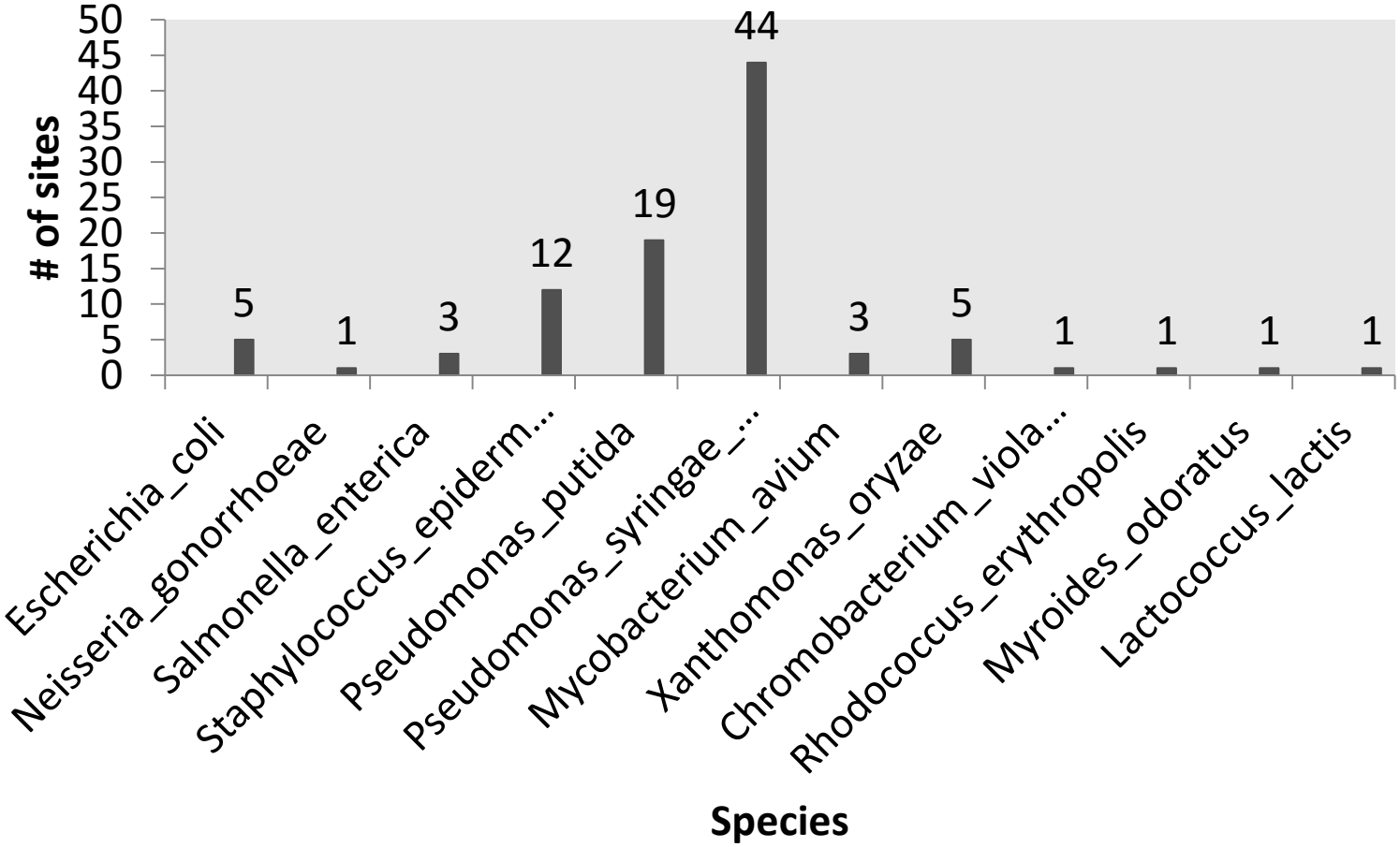


Fig. 3: Microbiome composition from November 2014 sample collection. *Pseudomonas* and *staphylococcus* species emerge as major contributors to baseline microbiome.

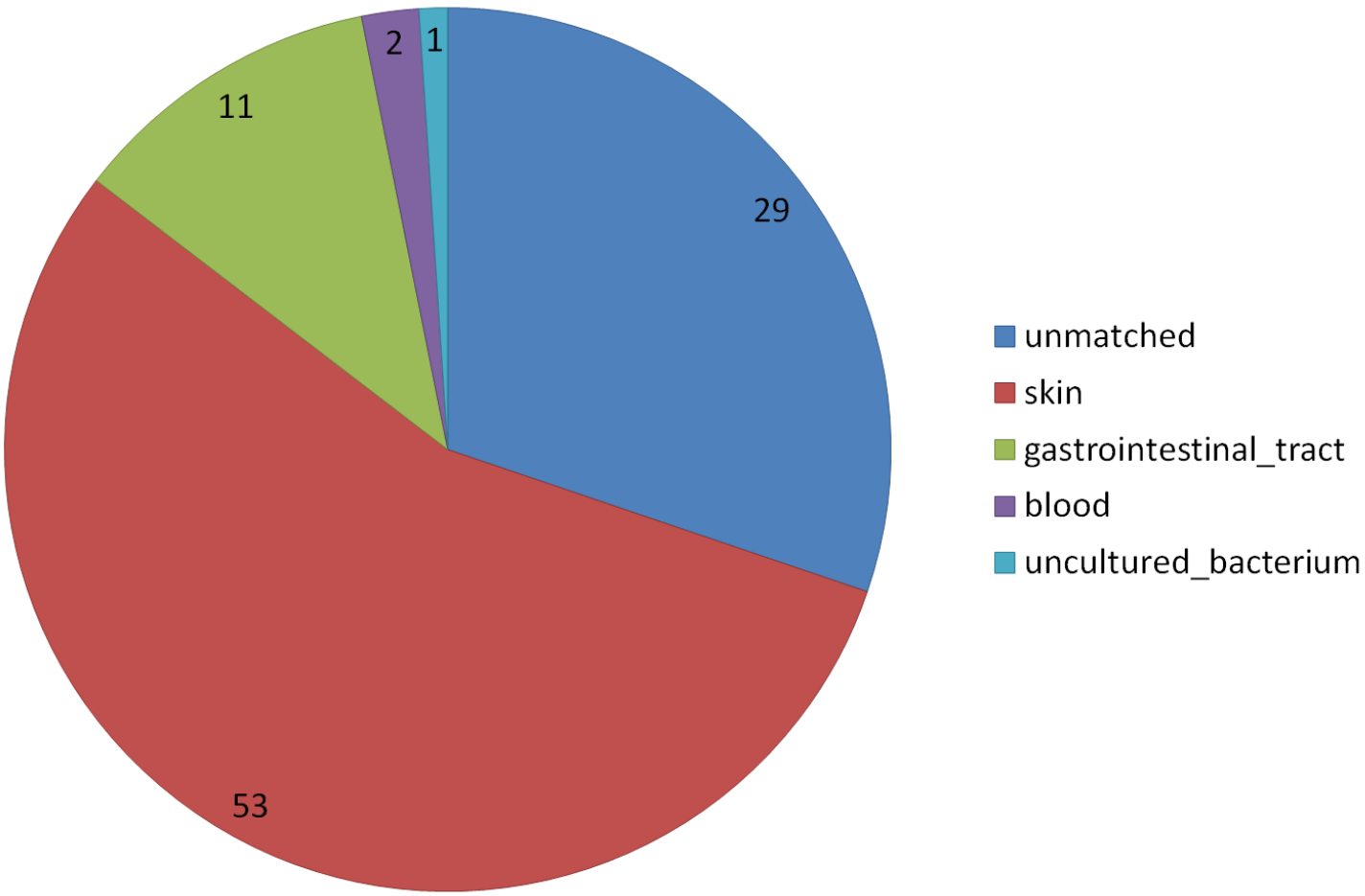


Fig. 4: Bacteria associated with body region for August 2014 sampling sites.

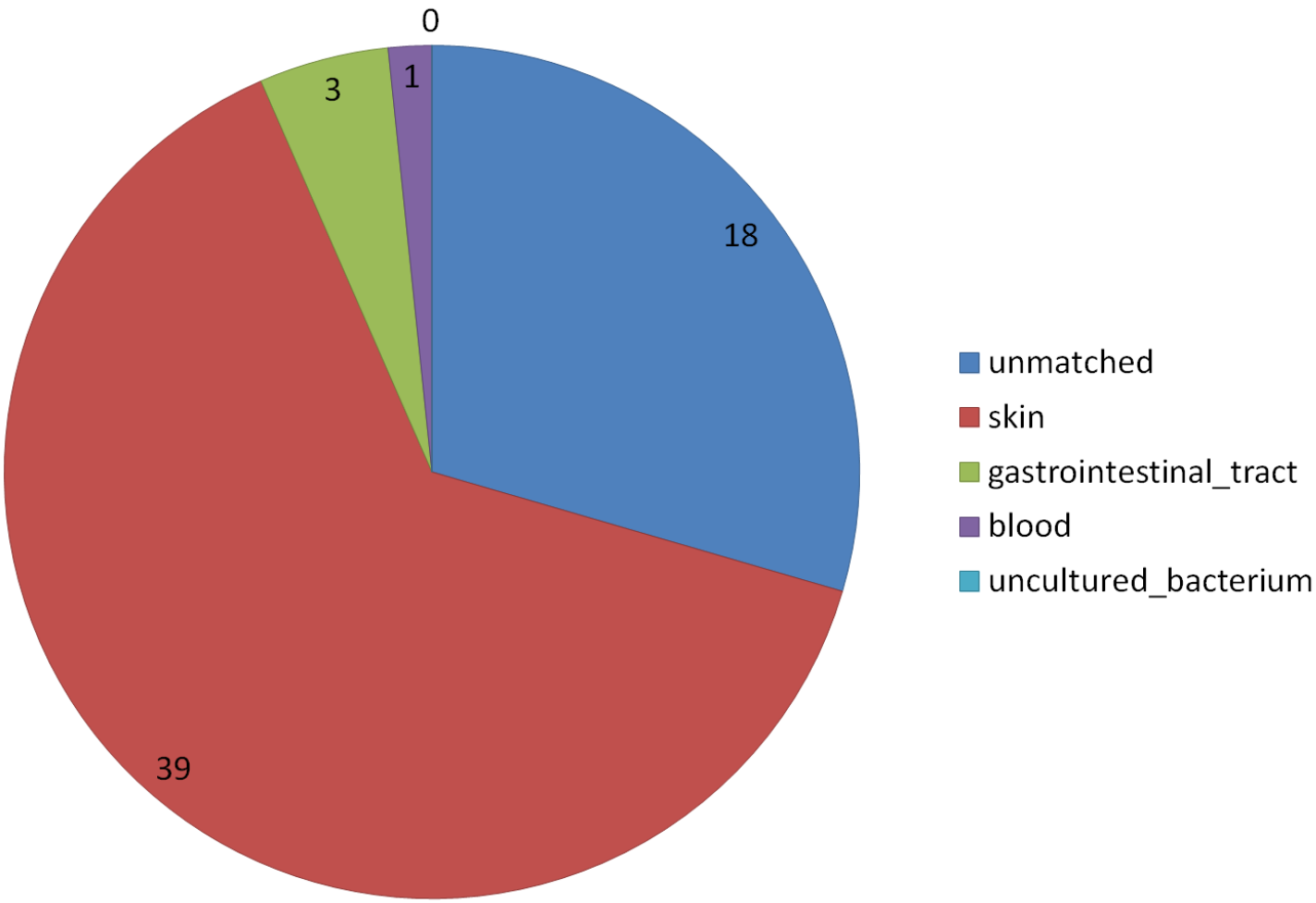


Fig. 5: Bacteria associated with body region for September 2014 sampling sites.

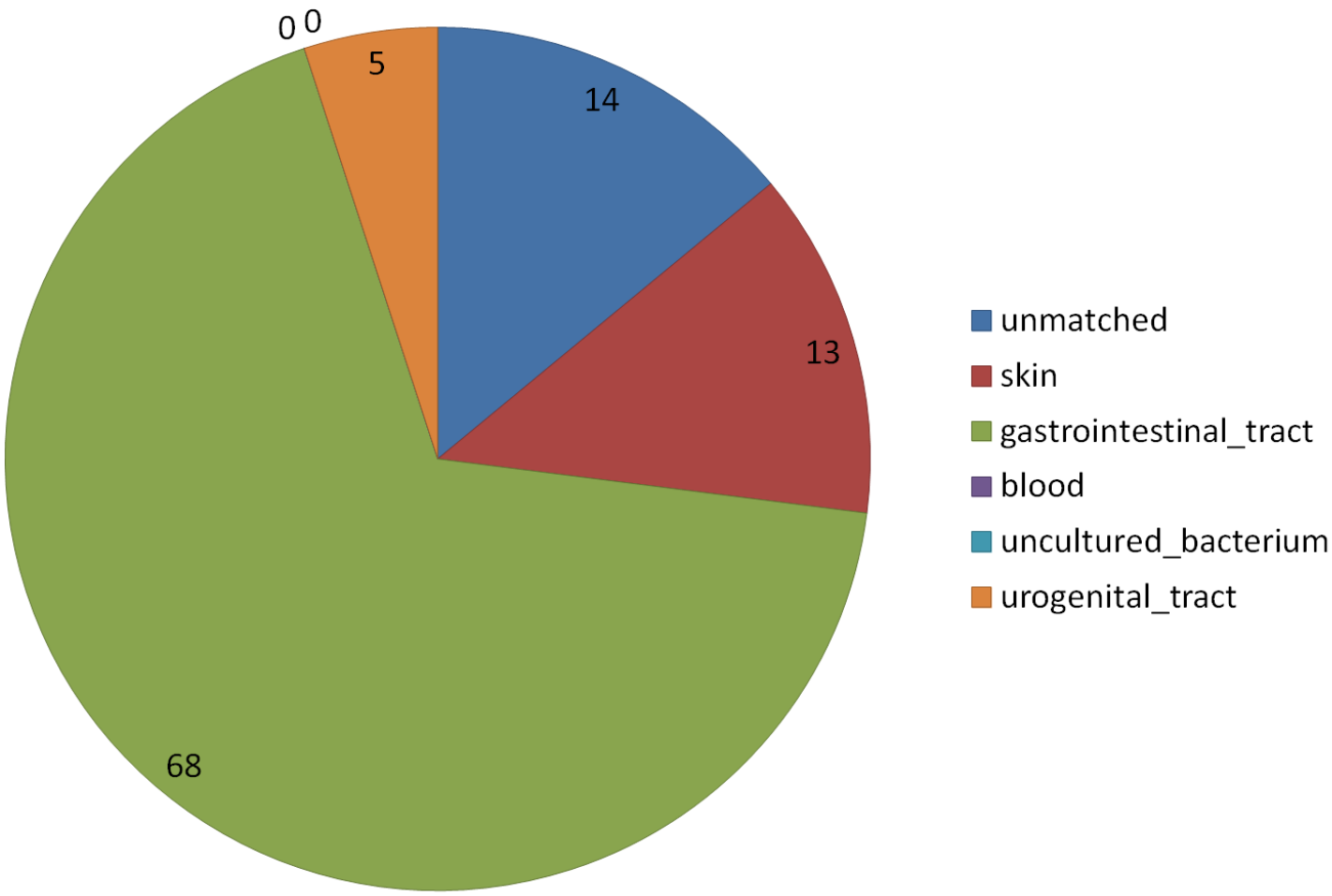


Fig. 6. Bacteria associated with body region for November 2014 sampling sites.

Virulence marker analysis

Sample	Pre-read capping	Post-read capping
SL124644	*Y. pestis* plasmid S. enterica serovar Choleraesuis str. SC-B67 (complete genome) C. diphtheriae plasmid pNG2 methylase gene	None
SL124645	M. tuberculosis H37Rv MysA & MysB genes (complete) *Y. pestis* tRNA-Asn gene (complete sequence) HPV type 23 (complete genome) Mycobacterium leprae rpoT gene Haemophilus influenza Rd KW20 (complete genome)	*Y. pestis* Angola, complete genome *Y. pestis* Antiqua, complete genome *Y. pestis* CO92 complete genome *Y. pestis* D106004, complete genome *Y. pestis* D182038, complete genome *Y. pestis* KIM, complete genome *Y. pestis* Nepal516, complete genome *Y. pestis* Pestoides F, complete genome *Y. pestis* Z176003, complete genome *Y. pestis* biovar Medievalis str. Harbin 35, complete genome *Y. pestis* biovar Microtus str. 91001, complete genome
SL124655	Klebsiella pneumoniae plasmid pRYCE21 transposon *Y. pestis* plasmid pIP1203 S. typhi outer membrane protein (ompC) gene A. baumanii	Klebsiella pneumoniae insertion sequence IS5 TnpA (tnpA) gene, complete cds; and disrupted ompK-36 gene, complete sequence Klebsiella pneumoniae p9701 plasmid Klebsiella pneumoniae plasmid pRYCE21 transposon Tn1000-like TnpA gene, partial cds Klebsiella pneumoniae putative RecF gene, partial cds Klebsiella pneumoniae strain H224 insertion sequence IS5, complete sequence Klebsiella pneumoniae strain KF3 plasmid pKF3-140, complete sequence Klebsiella pneumoniae strain KP1861/05 insertion sequence IS5-like ompK36 pseudogene, partial sequence Klebsiella pneumoniae strain KP300/08 insertion sequence IS4, complete sequence Klebsiella sp. KCL-2 plasmid pMGD2, complete sequence Klebsiella pneumoniae strain YMC 08/1/8 disrupted OmpK36 (ompK36) gene, partial sequence Klebsiella pneumoniae subsp. pneumoniae MGH 78578 plasmid pKPN3, complete sequence Klebsiella pneumoniae subsp. pneumoniae MGH 78578, complete sequence Klebsiella pneumoniae strain KP54/08 insertion sequence IS5, complete sequence Klebsiella pneumoniae strain NK29 plasmid pK29, complete sequence Klebsiella pneumoniae subsp. pneumoniae MGH 78578, complete sequence Klebsiella sp. KCL-2 plasmid pMGD2, complete sequence

Table 1. Summary of virulence marker analysis pre- and post-read capping for samples SL124644, SL124645, and SL124655.

DISCUSSION

One problem that surfaced from the original PathoMap study in New York City was that traces of *Y. pestis*, the bacteria known for causing the bubonic plague, were found in samples across the NYC subway system (Afshinnekoo et al.). This caused uproar and panic from city public health officials, not to mention many public news sources such as the New York Times and National Geographic. The problem with the PathoMap project is that the Mason Lab published any and all reads that were associated with a potential pathogen, without checking to make sure these reads provide appropriate coverage of the pathogen genome.

We developed a method called read-capping to help determine real hits to pathogens in the sample versus false hits. Table 1 shows a comparison of virulence marker analysis pre- and post-read capping for two samples. The first sample shows hits to potential pathogens upon first analysis, but after read-capping, these markers are no longer present in the sample. The second sample, however, displays markers for *Y. pestis* both pre- and post-read capping, indicating that this could be a potential pathogen present in our sample. Further validation would need to be performed to confirm these results (Derrien et al.).

Initial virulence marker analysis indicated that sample SL124644 had hits for a *Y. pestis* plasmid, the entire *S. enterica* serovar Choleraesuis str. SC-B67 genome, and the pNG2 methylase gene from *C. diptheriae*. The pNG2 methylase gene has been indicated in erythromycin-resistance in a 9500-kDa plasmid that was previously isolated. After capping virulence hits to 10 or greater reads, there were no reads mapped to the pNG2 methylase gene, *Y. pestis* plasmid, or *S. enterica* in this sample. Although our initial metagenomic analysis

identified reads with similarity to *Y. pestis* sequences, there is minimal coverage to the genome of these organisms, and there is not enough evidence to suggest these organisms are in fact present in this sample. For sample SL124645 there were over 10 validated virulence hits for *Y. pestis*, including some complete genome matches. This same sample showed initial matches to the *Mycobacterium leprae* rpoT gene, but after capping reads at over 10 reads, these were not validated. Initial analysis indicated hits to *Y. pestis* plasmid pIP1203 in sample SL124655 but further read capping indicated that this organism is not present in the sample. It did, however, confirm initial presence of *Klebsiella pneumoniae*.

Even if there are DNA fragments correctly matched to the right species, and there is evidence that it is coming from a living organism, this still does not mean the organism is a pathogen. Pathogenicity depends on many factors: infective dosage, immune state of the hosts, route of transmission, other competitive species, informatics approaches to species identification, horizontal transfer, bacterial methylome state and unique base modifications, and many others (Mason et al.). Most importantly, to be a pathogen it has to be shown to be infectious, usually along the lines of Koch's postulates.

In the case of the PathoMap study, the Mason Lab had data that indicated some species, but to say that they have knowledge about their source, or whether they are living organisms, would take more work. Most importantly, to claim wisdom that they know the best practice for large-scale city management is also beyond the scope of the PathoMap paper – it was the first survey of a city's transit system and serves as a snapshot. The really interesting data will come when we examine it over time, and compare it to other cities. The large amount of interest, excitement, and fear in the PathoMap study on city-scale metagenomics and

species identification has led to some questions about the data, which has been further clarified from the Mason Lab in an addendum. The Mason Lab does not claim that any of these organisms are alive, and most especially they have no evidence they are pathogenic.

Based on the results of Stulberg et al.'s study, there are fewer activities in tool and resource development than might be expected, considering that all of the participating organizations noted a priority need for new tools, technologies, and databases as foundational resources for the field. We should be cautious to publish data with remarkable hits to virulent strains that lead to public health panic. Establish validation standards within the field and continue to improve and build robust pipelines with validation. We need to admit the weaknesses and biases of our current tools and approaches to categorizing pathogenic metagenomic data. Going forward, the transparency of the methods, annotations, algorithms, and techniques has never been more essential.

ACKNOWLEDGEMENTS

I would like to thank James Clemens High School and the AP Biology classes for their partnership with this project. I would also like to thank the Educational Outreach team at HudsonAlpha for coordinating the sampling of JCHS. I want to acknowledge the Genomic Services Lab team for their assistance with sequencing and data analysis. This research was also presented at the 2015 Southeastern Medical Scientist Symposium (SEMSS) with the assistance of an undergraduate student travel award funded by an R15 grant from the National Institutes of Health/National Institute of General Medical Sciences (NIH/NIGMS).

REFERENCES

Afshinnekoo et al., Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics, CELS (2015), http://dx.doi.org/10.1016/j.cels.2015.01.001

Amies CR. A modified formula for the preparation of Stuart's medium. Canadian Journal of Public Health, July 1967. Vol. 58, 296 – 300

Board on Life Sciences, & Division on Earth and Life Studies (2007). The new science of Metagenomics: Revealing the secrets of our microbial planet. United States: National Academies Press.

Derrien, T., Estellé, J., Marco Sola, S., Knowles, D. G., Raineri, E., Guigó, R., & Ribeca, P. (2012). Fast computation and applications of genome Mappability. PLoS ONE, 7(1), e30377. doi:10.1371/journal.pone.0030377

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., … Nelson, K. E. (2006). Metagenomic analysis of the human distal gut Microbiome. Science, 312(5778), 1355–1359. doi:10.1126/science.1124234

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. Genome Res. 17, 377–386.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589–595.

Mason, C. (2015, February 17). The long road from data to wisdom, and from DNA to Pathogen. Retrieved April 19, 2016, from http://microbe.net/2015/02/17/the-long-road-from-data-to-wisdom-and-from-dna-to-pathogen/

Naccache, S.N., Federman, S., Veeraraghavan, N., Zaharia, M., Lee, D., Samayoa, E., Bouquet, J., Greninger, A.L., Luk, K.C., Enge, B., et al. (2014). A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res. 24, 1180–1192.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al.; MetaHIT Consortium (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464, 59–65.

Stulberg, E. (2015). An assessment of US microbiome research. Nature Microbiology, 1, 15015. doi:10.1038/nmicrobiol.2015.15

Wolfsberg, T.G., and Madden, T.L. (2001). Sequence similarity searching using the BLAST family of programs. Curr. Protoc. Mol. Biol. Chapter 19, 3.